INTRODUCTION

Mental health measurement, particularly of mood disorders, is notoriously difficult. This difficulty is largely due to challenges of robustly quantifying an individual's mental health status and ambiguity around tangible thresholds. In combination with primary care access issues, this results in the substantial under-diagnosis of mental health disorders with as many as 45-85% of depressed individuals estimated to never receive diagnoses of depression (Charon Gwynn et al., 2008; Verheij, 1996). As clinical outcomes under-report psychological morbidity, particularly for 'milder' cases of depression (Garrard et al., 1998; Lecrubier, 2007), quantitative social scientists commonly estimate psychological morbidity from questionnaires distributed to a sample of the population.

The 12-Item General Health Questionnaire (GHQ-12) is one of the most widely used responses in quantitative social science and epidemiology for the analysis of mental health trends. It contains 12 questions, each with 4 Likert response categories, which are conventionally summed to give a single score lying on a notionally meaningful singular dimension (Goldberg and Hillier, 1979). The popularity of the GHQ-12 is largely due to its ease of use, breadth of distribution, and capacity to reproduce "remarkably robust" results contrasted with longer initial versions (Goldberg et al., 1997). Initially focusing on diagnostic purposes for specifically at-risk individuals, the GHQ-12 has since been repurposed and validated across multiple languages and countries as a population screening tool for depression and depressive symptoms (Creed and Evans, 2002; Hankins, 2008a; Pevalin, 2000; Smith et al., 2010). The UK is no exception, where the GHQ-12 has become the canonical mental health measure for UK-based population studies (Propper et al., 2005; Thomson and Katikireddi, 2018; Weich et al., 2003), largely due to its inclusion as principal mental health outcome for a series of popular large-

scale surveys including the UK Household Longitudinal Survey (McFall, 2011) and the Health Survey for England (Mindell et al., 2012).

Despite being extensively validated using other leading mental health metrics, and longer versions of the same GHQ metric, there is still considerable debate as to what the GHQ-12 truly captures (Werneke et al., 2000; Ye, 2009). Due to its repurposed nature, different disciplines treat the GHQ-12 differently, treating it either as an indicator of cases of ill-health (Goldberg and Bridges, 1987; Thomson and Vittal, 2018), or as an indicator of population trends in mental health spectra (Hu et al., 2007; Weich et al., 2003). Researchers interested in underpinning processes of mental health, rather than population indicators, have extensively interrogated what is truly captured by the GHQ-12. This has led to numerous proposed factor structures with little consensus on whether more complex structures are truly adding value or if they result from over-interpretation of substantively meaningless, stochastic variation (Aguado et al., 2012; Hankins, 2008a; Stochl et al., 2016).

The critical importance of understanding composite, complex measures such as the GHQ-12 is clearly evidenced in genetic and psychological research., More strictly defined cases of depression produce higher estimates of genetic heritability for major depressive disorder (Cai et al., 2018). Similarly, mental illness and wellbeing, often used interchangeably in policy, can have widely differing covariates (Patalay and Fitzsimons, 2016; Westerhof and Keyes, 2010). This difference has been recognised both nationally and internationally by the UK Chief Medical Officer signposting greater understanding of positive and negative components of mental health as a public health priority (Davies, 2013, 2018). Obtaining a more comprehensive understanding of the underpinning, population-scale processes of mental health could offer more nuanced insight into these phenomena.

This paper uses novel Bayesian Exploratory Structural Equation Modelling techniques to estimate underpinning processes governing responses to the GHQ-12 in Wave 1 of Understanding Society. The estimation of latent scores facilitates more nuanced inference about processes underpinning mental health and provides an empirical framework to test the stability of these complex processes across populations. The increasing availability of Bayesian capabilities in latent variable software packages (Mplus, (Muthen and Muthen, 2017), blavaan, Merkle & Rosseel, 2018) enable the generation of interpretable posterior factor scores, which can be readily incorporated as outcomes in subsequent analyses. Taken together we argue Bayesian ESEM offers an opportunity for quantitative social scientists to extract more of existing mental health data. Furthermore, this method allows the relaxation of previously necessary methodological constraints which have been demonstrated to bias results towards simpler interpretations (Marsh et al., 2014). The factor structure provided by this procedure is subsequently evaluated against leading interpretations of GHQ-12 dimensionality. Moreover, all models are estimated using both Weighted Least Squares and Bayesian estimation to highlight the capacity of the latter for more flexible estimation of non-normal values, giving less biased results for potentially skewed factor variances and more explicitly discern preferential specifications (Muthén and Asparouhov, 2010).

The rest of the paper is organised as follows. Firstly, an overview of the methodological assumptions underpinning traditional factor analysis methods is provided. Secondly, we review previous GHQ-12 structures suggested from the literature, identifying common dimensions across studies. We then consider the data to be analysed from Understanding Society and outline a recently developed analytical approach, ESEM, which presents a less constrained approach to instrument decomposition than traditional factor analysis allowing for more nuanced factor structure. This analytical approach is then shown to identify a more theoretically informative factor structure for the study population, which is then contrasted with previous

structures. We demonstrate further flexibility by deploying the ESEM measurement invariance model taxonomy proposed by Marsh et al. (2009) with respect to gender. An extended conclusion section concludes by exploring the implications and limitations of this newly found structure, with recommendations for best practice for quantitative social scientists interested in more fully exploiting complex mental health outcomes to generate substantively meaningful insight.

1.1 Understanding Mental Health Measurement

The desire for population screening metrics can be situated as symptomatic of a wider cultural shift in mental health perception away from an "absence of illness" perspective. Where early studies of the ecology of mental illness focused on diagnoses of psychosis (e.g. Faris and Dunham, 1939), evidencing change via reduction in illness, contemporary approaches advocate a more holistic approach aiming to measure improvements in "wellbeing" (World Health Organization, 2013). This necessarily involves the conceptualisation of a mental health "spectrum", beyond binary caseness. Whilst complexity is increasingly recognised as an intrinsic part of mental health measures (Gnambs and Staufenbiel, 2018; Hu et al., 2007), quantitative social science commonly takes the reductive view of reducing questionnaire responses to binary "cases" or unidimensional constructs without adequate consideration of the methodological implications of doing so.

The underpinning theoretical assumption of summed GHQ-12 scores is that mental health lies on a single spectrum, thus all variation occurs along that spectrum. This unidimensionality is of critical relevance in literature which recommends or posits thresholds for the GHQ-12 above which an individual is considered at risk, or an "ill case" (Baksheev et al., 2011; Goldberg et al., 1998; Tait et al., 2003). In this literature, for caseness to be meaningful at the population level it is critical that a unit increase can be assumed to represent the same change in mental

status across individuals, an assumption that necessarily posits substantive equivalence in unit response change both *between* items and *within* items. Ultimately all research treating the GHQ-12 as unidimensional, whether continuous or categorical, necessarily posits that a unit-increase in summed GHQ-12 score implies the same change wherever it occurs across the metric (Brodersen et al., 2007; Marsh and Bailey, 1991).

Moving beyond internal validity, the context-sensitivity of GHQ-12 interpretation has been evident for some time (Werneke et al., 2000). Goldberg et al. (1998) noted the context-specific nature of thresholds in the GHQ-12, documenting the clear differences in specificity and discriminant capacity between-items across-countries. Despite this early acknowledgement, the issue has been largely disregarded in subsequent literature due to necessary impositions of overly simplistic analyses and validation measures.

Numerous studies have attempted to characterise the internal consistency of the GHQ-12 using relatively simple tests of Cronbach's Alpha (Cronbach, 1951) or by conducting Confirmatory Factor Analysis (CFA) (Jöreskog, 1969). Of the two, CFA is considered the superior approach, providing robust evaluations of capacity for model replication (Dunn et al., 2014). However, it still requires a refinement of dimensionality to "simple structure", implying zero cross-loadings across items (Asparouhov and Muthén, 2009a; Marsh et al., 2014). This means that each constituent item can only be empirically related to one underpinning construct, which has in turn been argued to lead to a reliance on overfitting models via model modification indices (Asparouhov et al., 2015), rendering the ostensibly *confirmatory* analysis theoretically *exploratory* (Fabrigar et al., 1999; Schmitt, 2011). The issues surrounding this approach to social science metrics are not solely empirical, there are also substantive concerns when using traditional CFA approaches (Conway and Huffcutt, 2003). The most pressing of these concerns is the orthogonality imposed by simple structure, which is particularly unrealistic for the GHQ-

12, as “nonzero cross-loadings are inherent in psychological measurement” (Marsh *et al.*, 2014, pp.88).

1.2 Proposed Factor Structures

Whilst longer versions of the GHQ have long been accepted to be multidimensional (Graetz, 1991; Martin, 1999), consensus is less forthcoming for the GHQ-12. Of the variety of proposed interpretive structures for the GHQ-12, there are three common interpretations. The simplest treats the GHQ-12 score as a unidimensional construct, taking the summed scores as a response, occasionally with an adjustment for positive and negative phrasing. This approach is backed up by a large body of research, which uses CFA to conclude that the unidimensional interpretation of the GHQ-12 is the most compelling (e.g. Aguado et al., 2012; French and Tait, 2004; Winefield et al., 1989). Despite this research concluding in support of unidimensionality there is another important consideration here.

Studies commonly cite high correlations between modelled multidimensional factors as justification for unidimensional interpretation (e.g. Gao *et al.*, 2004; Gouveia *et al.*, 2010; Padrón *et al.*, 2012; Fernandes and Vasconcelos-Raposo, 2013; Romppel *et al.*, 2013). Recent work using simulated data however has demonstrated that imposition of simple structure can artificially inflate correlations between modelled factors (Asparouhov and Muthén, 2009a; Marsh et al., 2014). Thus whilst it is not uncommon for reported correlations to be greater than 0.9 (e.g. Aguado et al., 2012; Campbell and Knowles, 2007) or 0.95 (e.g. Sweeting et al., 2009; Wang and Lin, 2011), taking these correlations as justification for unidimensionality risks a self-fulfilling prophecy of simplicity begetting simplicity.

Whilst many studies call for unidimensional interpretation there are numerous proposed multidimensional GHQ-12 structures. Several two-factor solutions have been proposed and

validated using CFA techniques. These two factors most commonly involve a "Depression/Anxiety" construct and a "Social Dysfunction" construct (Andrich and Schoubroeck, 1989) (given by GHQ-12 items 2, 5, 6, 9, 10 and 11, and 1, 3, 4, 7, 8 and 12 respectively - see Supplementary Material for full item list). "Depression/Anxiety" relates to the emotional component of psychological distress, whereas "Social Dysfunction" relates to the social functioning of the distressed individual. This structure, albeit with different labelling, has at various points been identified as the best fit to data from the UK (Smith et al. 2010), New Zealand (Kalliath et al., 2004), Brazil (Gouveia et al., 2010), Japan (Suzuki et al., 2011), Germany (Schmitz et al., 1999), Italy (Politi et al., 1994) and Turkey (Kiliç et al., 1997). Systematic reviews and meta analyses also consistently identify these two factors most commonly in both two- and three-factor solutions (Gnambs and Staufenbiel, 2018; Picardi et al., 2001; Werneke et al., 2000). It is important to note that these groupings align with the positive or negative phrasing of the constituent items. "Social Dysfunction" items are all positively worded and "Depression/Anxiety" items are all negatively worded, which has invited debate as to whether this structure solely reflects differences in phrasing (Hankins, 2008b; Stochl et al., 2016).

Increasing multidimensional complexity brings consideration of three-factor solutions, which most commonly identify "Social Dysfunction" and "Depression/Anxiety" constructs alongside a third construct referring to some variant of "Loss of Confidence". This structure was initially identified by Worsley and Gribbin (1977) in the first factor analysis of the GHQ-12. They found a third "Loss of Confidence" construct alongside "Social Performance" and "Anhedonia-Sleep Disturbance" (approximating Social Dysfunction and Depression/Anxiety respectively). For Worsley and Gribbin this third posited dimension was constructed of four items (6, 9, 10 and 11) pertaining to losing confidence or feeling worthless or depressed, evidence of which has been subsequently supported (Campbell et al., 2003; Penninkilampi-Kerola et al., 2006;

Vanheule and Bogaerts, 2005). Notably, Worsley and Gribbin did not refine to a simple structure solution, as there were no structures to test against, so this solution contains cross loadings.

A similar structure was published by Graetz in 1991. Graetz analysed a large scale (N= 8998) Australian sample, resulting in a three-factor structure comprising "Anxiety/Depression", "Social Dysfunction" and "Loss of Confidence". The Loss of Confidence factor is similar to that of Worsley and Gribbin – but presented in the simple structure format of CFA, with non-zero cross-loadings eliminated. This 3-factor structure has been the most widely accepted multi-dimensional structure since its introduction, supported by subsequent work both in the UK (e.g. Cheung, 2002; Martin& Newell, 2005; Shevlin & Adamson, 2005) and overseas (Gnambs and Staufenbiel, 2018; Padrón et al., 2012).

Whilst several proposed structures from the literature are strongly validated, there remain certain issues with their research design. Firstly, with the exception of the Graetz 3-factor model (Graetz, 1991), structures are commonly generated and validated using data with small sample size and limited contextual scope. It is not uncommon for validation studies to report sample sizes under 500 (e.g. Khan et al., 2013; Martin and Newell, 2005). Moreover, studies with greater statistical power commonly draw samples from heavily context-specific populations such as primary care users (e.g. Werneke *et al.*, 2000) or high-school students (e.g. Suzuki *et al.*, 2011), and not the general population. Secondly, there is little methodological consensus on effective measures and fit criteria for recommending any given structure over another, despite efforts to establish consistent approaches (Conway and Huffcutt, 2003; Hu and Bentler, 1999; Marsh et al., 2010a).

There is clear substantive interest to social researchers in moving beyond understanding observed, aggregated measures of complex phenomena such as mental health. It is of critical

importance to develop a better understanding of the underpinning processes driving these measures. Empirically, this means partitioning response variance into common and unique variation across items, explicitly factoring for individual variability in response (Conway and Huffcutt, 2003). There is clearly a wealth of data available for the GHQ-12 in the UK, collected alongside rich demographic and spatial information in several panel studies (e.g. Gnambs and Staufenbiel, 2018). In light of methodological developments in the modelling of complex responses, this data needs comprehensively re-evaluating to ensure it is fully exploited for given populations. These complex methodologies readily output individual scores on the proposed dimensions, which can be used to triangulate evidence around more complex aspects of mental health.

This study proposes a multidimensional series of latent factors underlying GHQ-12 responses. We use the ESEM framework to test for measurement invariance, finding evidence of strict measurement invariance with respect to gender. The interpretation of multidimensional latent structures allows insight into underpinning processes producing the observational data. We argue that Bayesian ESEM offers a readily available methodology with which the increasing number of social scientists interested in mental health can ask and answer more realistically nuanced questions of mental health metrics. Furthermore we clarify and make explicit the empirical and substantive benefits of doing so in the quantitative analysis of a subject as nuanced and individually heterogeneous as mental health.

2 *DATA AND METHOD*

2.1 Data

This study uses the first wave of the Understanding Society Survey (US), a nationally representative annual UK panel survey . Data collection for Wave 1 took place between January 2009 and March 2011 and involved the completion of full interviews by 47,732 respondents (McFall, 2011). The GHQ-12 was administered via face-to-face interview and completed by 40,452 respondents. For a comprehensive overview of Understanding Society Data Collection and methods see Quality Profile (Lynn and Knies, 2016).

Observations were dropped from the original dataset if they did not record any responses to GHQ-12 items. Of the 40452 who provided responses to any GHQ-12 items, 39700 responded fully. Estimation was carried out on the 40452 respondents, as Bayesian estimation does not require the strict missing completely at random assumption typically required for modelling of ordinal factor indicators (Muthén and Asparouhov, 2012). Table 1 gives the response characteristics of the survey sample, demonstrating that including partial respondents does not appreciably change demographic characteristics other than to allow us to increase the number of non-whites considered in the sample. For initial exploration sample weights were taken and applied from Wave 1. Sample weights were applied to Weighted Least Squares but not Bayesian analyses, as weights are not supported in Bayesian estimation.

[*Table 1 about here*]

There is strong evidence of different response patterning across positive and negative GHQ-12 items, shown in Figure 1. There are clearly different modal responses for positive and negative items, giving rise to the common factor structures reflecting positive and negative items.

Despite strong patterning in responses being somewhat corrected for in binary interpretation, Likert scores have been demonstrated to allow greater discrimination between different dimensional structures (Campbell and Knowles, 2007; Smith et al., 2013). For this reason we use full information Likert scoring here (see supplementary material for response coding).

[Figure 1 about here]

2.2 Methodology

To explore the GHQ-12 data a novel approach termed Exploratory Structural Equation Modelling (ESEM) is deployed (Asparouhov and Muthén, 2009). ESEM combines the best elements of restrictive CFA and unstructured exploratory factor analysis (EFA – which is actually a special case of ESEM, Asparouhov and Muthén, 2009b). The key contribution of ESEM for this research is the specification of non-zero cross-loadings on constituent items (Asparouhov and Muthén, 2009). As outlined above, in a literature which cites high factor correlations as justification for unidimensional interpretation, it is especially important to guard against inflated estimates of factor correlations resulting from imposing zero cross-loadings (Asparouhov et al., 2015; Marsh et al., 2010b, 2014). Beyond the empirical, there is also a substantive argument against refining to simple structure when analysing complex social outcomes such as psychological constructs. Where multiple and inter-related underpinning processes are likely to give rise to any specific item response nonzero cross-loadings ought not be viewed an aberration but a logically anticipated representation of complex, underpinning constructs (Marsh *et al.*, 2014). All analyses are carried out within the Mplus software environment (Muthén and Muthén, 2018).

The model is estimated using both Bayesian and Mean and Variance Adjusted Weighted Least Squares (WLSMV) estimation to demonstrate the discriminant superiority of the former. A series of EFA solutions are estimated in which each and every item loads on each of 2-5 constructs from the 40452 Likert GHQ-12 responses. All solutions are rotated using Geomin rotation ($\varepsilon = 0.01$), which has been demonstrated to be optimal when little is known about the true underlying structure, and when suspected variable complexity is greater than one – meaning there is an expectation that there will be cross-loadings (Browne, 2001; McDonald, 2005). As we are conducting a thematically exploratory analysis, we are interested in inferring

maximal information about unknown latent dimensions emergent from the analysis, thus retention is dictated by a cutoff value of 0.224 as it represents 5% explained variance of the latent dimension. As the need to exclude loadings purely for methodological reasons has been relaxed, this value is a far more tolerant exclusion criterion than typically advocated in the literature. To test this criterion, models were rerun with all loadings present but with informative prior distributions of N(0, 0.01) for these small EFA loadings in Bayesian analysis, and cross loadings set to 0 in target rotation for WLSMV estimation. If estimated loadings under these constraints were under 0.1 (<1% of factor variance explained) then this was taken as convergent evidence supporting zero cross-loadings. Factor loadings above this were replaced back into the structure.

Partial respondent inclusion has an impact that is not consistent across methods. Under Bayesian estimation the posterior distribution is considered asymptotically consistent with full information maximum likelihood (FIML) so is unbiased in the face of missingness patterned with respect to known observations and known covariates (Asparouhov and Muthén, 2010). Conversely, although considered the optimal estimation for ordinal factor indicators (Schmitt, 2011), WLSMV essentially estimates using pairwise deletion under missingness, which is asymptotically consistent with FIML only when data can be assumed to be missing completely at random. As WLSMV estimates are provided primarily to support and corroborate findings from the Bayesian solution, which can incorporate the increased non-white population included in the partial respondents, partial respondents are included for all models.

Models are parameterised as multinomial probit regressions of each item on the underpinning latent factor (Asparouhov and Muthén, 2010; Muthen and Muthen, 2017, pp.62)). Factor loadings can be interpreted as likelihood changes in the log-odds of changing response category on the response variable, illustrating the strength of the relation between the underpinning

dimension and the probability of response change in the associated item (Chen, 2007). Bayesian estimation allows less biased estimation of non-normal parameters, such as variances, than WLSMV estimates. It also does not require the assumption of normality of the parameter estimates, with prior variance-covariance estimates instead drawn from an inverse-Wishart distribution (Asparouhov and Muthén, 2010).

To further demonstrate the utility of the ESEM approach, we subsequently tested the invariance of the proposed structure with respect to gender using the model invariance taxonomy developed by Marsh et al. (2009). Comparison of group differences for measurement invariance testing is not currently available in Bayesian estimation of ordinal responses in Mplus. Thus, whilst evidence for the factor structure is gathered using Bayesian estimation, it is not possible to evaluate the measurement invariance of this structure in a Bayesian framework. As such we do not report results for this fully here but provide model syntax, results and further references for emerging invariance methodologies in Supplementary Material.

Model fit is used to both evaluate the proposed structure and to ensure that it offers greater predictive capacity than structures available from previous literature. For Bayesian estimation, fit is evaluated solely via Posterior Predictive Checking (PPC) for which we present posterior predictive p-values and 95% credible intervals for each model. In Mplus PPC is an extension of the likelihood ratio statistic taken to be an indication of the model's capacity to reproduce the data and summarise the posterior distribution of the residuals (Asparouhov and Muthén, 2010). As such, the statistic is upward biased by sample size, rendering us unlikely to realistically observe a non-zero p-value, thus we consider reduction towards zero of the predictive credible interval as evidence of improvement in predictive capacity, indicating proximity to a plausibly zero-sum discrepancy function(Gelman et al., 1996; Marsh et al.,

2004). We also present mean absolute factor correlation values (taken from the Bayesian estimation) to evidence substantive dissimilarity between modelled constructs with lower values indicating greater discriminant validity. In order to gain traditional fit statistics to facilitate comparison across methodologies, such as the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) and the Standardised Root Mean Square Residual (SRMR), WLSMV estimation was also carried out. For a more comprehensive discussion of the definitions of these fit indices see Hu and Bentler (1999) and Supplementary Material. Evidence of good fit was taken from guidelines proposed by Bentler (2007) and Muthén and Muthén (2014).

3 RESULTS AND DISCUSSION

Initial analysis specified a series of EFAs with between 2 and 5 factors, the results of which are given in Table 2. In the frequentist estimation, goodness of fit statistics evidence better model improvement for each added dimension up to a five-factor solution. In the Bayesian estimation, it is more explicit that none of the specifications provide an adequate fit to reproduce the 40,452 individuals' mental health response, as shown by the consistent posterior p-values of 0.000, indicating almost no likelihood of fully recovering our data for any model. However, there is evidence of improvement of fit in the posterior credible interval with progression towards zero indicative of better fit. Predictive capacity improves with added factors up to the fourth factor; however, the five-factor solution is less robustly estimated. Reduction in the lower bound shows that a fifth factor could serve to improve predictive capacity but the increased upper bound suggests it is more probable that it will reduce the predictive capacity of the model. Therefore, the four-factor structure is carried forward as it offers the most robust reproduction of observed values.

[Table 2 about here]

Having identified that the four-factor solution was the most well supported from the estimation of the initial EFA whilst all cross-loadings are specified and estimated, the subsequent solution could benefit from refining to a comparable structure. This was implemented by re-specifying the model with all cross-loading values under 0.224 being omitted as outlined in Methods. Factor loadings can be interpreted as probit regression coefficients of each item on the unit-standardised latent factor and are comparable both within and between factors.

[Table 3 about here]

Table 3 gives the loadings for the four-factor ESEM solution. The four factors are labelled "Lowered Self Worth", "Social Dysfunction", "Stress" and "Emotional Coping". These were chosen to reflect existing names from the literature. The key difference from previous structures is that items 1, 2, 4, 5, 6, 7, 9 and 12 now load on multiple underlying dimensions. The loadings of Factors 1 and 2 retain the broad differentiation between positive and negatively worded items of previously proposed factors, and are named to reflect this, however, they differ empirically because they contain cross-loading items, but the. Higher individual factor scores across Factors 1 and 2 indicate higher levels of mental distress.

The third factor, here termed "Stress" is seen to load most strongly on items associated with feeling under strain and loss of sleep. It is very similar to a factor found in one of the earliest factor analyses of the GHQ-12 by Worsley and Gribbin (1977), differing only in that they found it also loaded on item 12.

The emergence of the fourth factor, termed "Emotional Coping" is a distinctive finding. Item 9 initially returned a loading of 0.203, and was specified a prior distribution of N~(0,0.01) in the simplified analysis, however it subsequently returned a value of 0.210 and thus could not be assumed zero given our criteria. Factor 4 is most notable for having both positive and negative loadings. It is *negatively* associated with Item 4 – "feeling capable of making decisions", but *positively* associated with feeling unhappy or depressed, not enjoying day-to-day activities and not feeling happy. As such individuals with high scores on Emotional Coping are those experiencing negative affect as captured by items 7, 9 and 12, whilst feeling capable of making decisions as captured by item 4, indicating a degree of (at least perceived) perseverance in the face of the distress. The negative loading highlights an ambiguity in interpreting latent variables. There is multidimensional symmetry between modal loadings in each latent axis. More simply, an empirically identical interpretation would be the inverse, with

negative loadings on Items 7, 9 and 12 and a positive loading on Item 4. This would in turn invert the poles of the underlying dimension, with higher scores indicating positive rather than negative outcomes. Emotional Coping is structurally most similar to the "Sleep Disturbance/Anhedonia" construct found in work by Worsley and Gribbin (1977) although with the key difference in the negative loading of Item 4.

Having detailed the theoretical implications of the proposed structure it is necessary to understand the substantive and empirical implications. To do so we evaluate the substantive dissimilarity between factor constructs .

[Table 4 about here]

Table 4 presents the modelled factor correlations. The factors with the highest correlation are Lowered Self Worth and Social Dysfunction, with a coefficient of 0.68. Although it is the highest identified in this structure, it is still low relative to that found in the existing literature. Stress is the most statistically dissimilar and therefore most substantively distinct factor, exhibiting uniformly low correlations with the other dimensions.

The absolute proportion of variation in one latent variable predictable solely from the other is given by the square of the correlation coefficient (Kish, 1954). For instance, knowing the modelled Lowered-Self-Worth scores for all individuals would only allow the prediction of 46.24% (0.4624) of the variation in Social Dysfunction scores, despite these factors having the highest modelled correlation of 0.68. This is even more notable for Stress and Social Dysfunction, with a predictive capacity of 1.2%, leaving 98.8% of variation unexplained.

To further demonstrate the utility of this framework the invariance of the structure was evaluated with respect to gender using the taxonomy developed by Marsh et al. (2009). It is reported only in Supplementary Material as currently it is not possible to estimate invariance

in categorical indicators using Bayesian estimation, however the structure evidenced strict measurement invariance. Having identified a parsimonious summary of data from our model it is important to connect back to the wider literature on mental health structures to establish that we could not have achieved a similarly well-fitting model solely using information that existed a priori. The literature has proposed many different structures, seen in Table 5, which gives context to seven exemplar proposed structures against which we evaluate the structure. The full specification of these models can be seen in Supplementary Material.

[Table 5 about here]

Table 6 presents fit statistics for all seven models alongside the proposed ESEM structure. The mean absolute factor correlation is also presented for each specification to evidence the dissimilarity of the factors estimated in the model structure. The four-factor solution provides the best fit across every measure of fit under both Bayesian and WLSMV estimation.

[Table 6 about here]

The best performing structure outside of the ESEM solution is the original Worsley and Gribbin (1977) structure with its non-zero cross-loadings. Factor structures which address the inter-dependencies of the items via error covariance or non-zero cross-loadings perform well across all the fit statistics, which presents an argument for the adoption of a more realistically complex specification of mental health. The benefit of cross-loadings is most clearly borne out in the mean absolute factor correlation, which is far lower for the four-factor structure than other solutions.

4 LIMITATIONS AND FUTURE RESEARCH

Whilst we are making the case for the capacity of novel methodologies coupled with large datasets to offer greater insight into complex outcomes it is important to highlight the limitations of this study. Whilst the structure proposed here provides the best fit for the data in Wave 1 of Understanding Society in the UK (McFall, 2011), it is necessary to test on a wider range of data beyond this spatiotemporally specific dataset. More research is required in order to understand fully what can be gained from the modelling of decomposed processes underpinning survey instruments. Firstly, the structure is still specific to the current dataset. It should also be noted that whilst tests of measurement invariance evidence internal consistency, they are also sample-specific (Supplementary Material). In the instance of external application of the structure, the demonstrated invariance would require re-evaluation. Moreover, whilst we evidenced strict measurement invariance with respect to gender, we did not test measurement invariance for further demographics in Table 1.As the invariance analysis was intended to be illustrative, and full information was unavailable for key demographics of interest we did not want to introduce further potential confounding by unobserved covariates influencing respondent tendency across constituent survey sections (Knies, 2017). Whilst it has been demonstrated that this structure provides a superior fit across the full population and that this holds consistent for males and females, measurement invariance should be investigated across different geographical and socio-demographic groups. There is considerable literature validating the overall GHQ-12 as a screening instrument over time and space (e.g. Gnambs and Staufenbiel, 2018), there is far less written on the temporal or geographical stability of its underpinning latent structure (Goldberg et al., 1997). This approach should be further applied across spatiotemporal contexts using the wealth of existing GHQ-12 data.

Secondly, whilst it is clear that fit indices favour the proposed structure over previous structures, it is important to reiterate that the purpose of this was to highlight the enhanced predictive capacity gained from constructing dimensions from the given dataset. Empirically we are asking more of the previous structures in providing an out of sample test, than the solution proposed here. As such it is important to establish the proposed structure here offers more substantively in terms of predictive capacity than simpler structures. Recommendations for fit indices come from studies using sample sizes far smaller than those used here (Bentler, 2007; Hu and Bentler, 1999). As such, in isolation there are multiple structures tested here which would be accepted as "adequate" under such criteria. Whilst the Bayesian fit statistics presented are more robust to sample size, more work is needed on the appropriateness of fit-statistic thresholds in the face of increasingly large sample sizes (Muthén and Asparouhov, 2010). Moreover, canonical reliability estimates are not yet available or optimised for Bayesian estimation of categorical factor indicators. For instance, whilst the TLI and RMSEA are considered appropriately robust to complexity (Marsh, 2009), the Bayesian PPC fit statistic presented here does not penalise as strongly for increasing numbers of parameters as other measures such as Deviance Information Criterion (Spiegelhalter et al., 2002), which are currently unavailable. Furthermore, methodological advances evolve rapidly (for instance see ongoing invariance work; Oberski, 2014; Seddig, 2018), and often optimal approaches are methodologically untenable due to software limitations, rather than theoretical implausibility. Against this backdrop, proposed structures must be evaluated not just numerically, but on a theoretical basis, identifying whether the more complex structure offers greater understanding. The mean absolute factor correlations give some indication of this in terms of substantive dissimilarity. However, comprehensive evaluation of this requires the investigation of the predictors of different constructs, demographically, geographically and socially to evidence whether they truly add to our understanding of different processes. As such, further work is

necessary taking decomposed constructs as responses with data for which the structure is validated.

Similarly, given the use of the GHQ-12 as an external validation instrument (e.g. Mukuria *et al.*, 2014), it is important to incorporate these more nuanced understandings in evaluating what is being captured by other existing or new metrics (e.g. Tennant *et al.*, 2007). Adopting this approach will allow researchers to inferentially attempt to understand similarities between underpinning mental health processes across populations. One such avenue is to investigate if any of the structures identified here are associated with similar underpinning processes of more recently developed well-being metrics such as the Short Warwick-Edinburgh Mental Well-Being Scale (Stewart-Brown et al., 2009). This would facilitate empirical contribution to the contested relationship of mental illness and well-being (Westerhof and Keyes, 2010).

There are further limitations in the interpretation of the resulting constructs. It has been suggested in GHQ-12 literature that multidimensional factor structures are simply a product of over-interpretation of spurious variance in negatively worded items (Hankins, 2008a). It is important to highlight the unidimensional correlated-error model performs very well, given its brevity. It seems reasonable to infer from this the potential for multidimensionality being solely the result of phrasing, only if one assumes unidirectional causality, i.e. items were grouped into positive and negative items at random, rather than based upon conceptually different measurements (Gnambs and Staufenbiel, 2018; Stochl et al., 2016). Empirically these two scenarios would present identically, although it seems reasonable to assume greater likelihood of the latter.

Replicating this analysis across different temporal and spatial contexts is a clear avenue of further research. We highlighted a framework under which these contexts can be tested for invariance and strongly advocate doing so. Whilst understanding the stability of the structure

across contexts is undoubtedly important, it is also important to demonstrate its worth in further developing understanding of what is being captured by these measures in large scale surveys. It is important, therefore, to deploy complex metrics by taking them as responses in spatial, social, structural and epidemiological studies. This is increasingly recognised in quantitative social science, evidenced by recent developments in genomic SEM (Grotzinger et al., 2019). It is clearly of greater benefit to social scientists in any analysis positing putative causal mechanisms to robustly link predictors with *underpinning processes* of mental health than with aggregate, unidimensional questionnaire responses.

5 CONCLUSIONS

The results of this study show that of the factor structures tested here the four-factor structure provides the best representation of GHQ-12 responses from the Understanding Society data, Wave 1. This is evidenced by traditional fit statistics as well as modelled factor correlations demonstrating greatest substantive dissimilarity. This involves the specification of two previously underexplored constructs, here termed "Stress" and "Emotional Coping". Emotional Coping is particularly notable as the presence of a negative loading highlights the capacity for underpinning constructs to mask the presentation of psychological distress in the aggregated metric. Furthermore, we find evidence of strict measurement invariance with respect to gender, although it is not currently possible to substantiate this for categorical latent variable indicators in a Bayesian framework.

Within the wider GHQ-12 literature there is little mention of dimensions analogous to the Stress and Emotional Coping structures beyond Worsley and Gribbin's early analysis (1977). Stress-related constructs are also proposed in a structure drawn from a Spanish population (Sánchez-López and Dresch, 2008), and notably the "thematic analysis" of GHQ-12 content

by Martin (1999). Constructs here would not have been identified using traditional CFA techniques as they consist largely of cross-loadings. The incorporation of cross-loadings considerably improved model fit, and the low modelled factor correlations suggest they are capturing substantively distinct processes. These low correlations in the results support suggestions that high correlations may be an artefact of restrictive modelling procedure thus we caution against refining to simpler dimensionality based purely on apparently high factor CFA correlations (Marsh et al., 2009).

The key message of this study is the capacity of large-scale datasets to contribute more comprehensive understandings of mental health outcomes in large, heterogeneous populations. Whilst this is not a new idea (Hu et al., 2007; Mukuria et al., 2014), the adoption of less stringent exclusion criteria in model selection and the incorporation of ESEM methodologies is something that is underexplored in large-scale survey analysis. The resultant individual posterior factor scores from such analyses are readily obtainable and offer inferential insight into processes which are more robust to confounding by item-specific variance and measurement error. In combination with large-scale surveys, decomposed metrics have the potential to offer a more comprehensive understanding of the similarities between different underpinning processes both between and within metrics.

In conclusion this study, despite the above limitations, is one of the first to combine Bayesian and ESEM methodology with large-scale survey data in the UK and has illustrated the inferential benefits of doing so for research into population-level mental health determinants. Further research is needed to validate these findings, using data from wider contexts, and contextualise them against differing mental health responses.

References

Aguado, J., Campbell, A., Ascaso, C., Navarro, P., Garcia-Esteve, L., Luciano, J. V, 2012. Examining the factor structure and discriminant validity of the 12-item General Health Questionnaire (GHQ-12) among Spanish postpartum women. Assessment 19, 517–25. https://doi.org/10.1177/1073191110388146

Andrich, D., Schoubroeck, L. Van, 1989. The General Health Questionnaire: a psychometric analysis using latent trait theory. Psychol. Med. 19, 469–485. https://doi.org/10.1017/S0033291700012502

Asparouhov, T., Muthén, B., 2010. Bayesian analysis using Mplus: Technical implementation. Los Angeles Muthén Muthén 1–38.

Asparouhov, T., Muthén, B., 2009a. Exploratory Structural Equation Modeling, Structural Equation Modeling: A Multidisciplinary Journal. https://doi.org/10.1080/10705510903008204

Asparouhov, T., Muthén, B., Morin, A.J.S., 2015. Bayesian Structural Equation Modeling With Cross-Loadings and Residual Covariances: Comments on Stromeyer et al. . J. Manag. 41, 1561–1577. https://doi.org/10.1177/0149206315591075

Asparouhov, T., Muthén, B.O., 2009b. Exploratory Structural Equation Modeling, Structural Equation Modeling: A Multidisciplinary Journal. https://doi.org/10.1080/10705510903008204

Baksheev, G.N., Robinson, J., Cosgrave, E.M., Baker, K., Yung, A.R., 2011. Validity of the 12-item General Health Questionnaire (GHQ-12) in detecting depressive and anxiety disorders among high school students. Psychiatry Res. 187, 291–6. https://doi.org/10.1016/j.psychres.2010.10.010

Bentler, P.M., 2007. On tests and indices for evaluating structural models. Pers. Individ. Dif. 42, 825–829. https://doi.org/10.1016/j.paid.2006.09.024

Brodersen, John, McKenna, S.P., Doward, L.C., Thorsen, Hanne, Morabia, A., Zhang, F., Wilson, J., Jungner, G., Mayor, S., Pashayan, N., Powles, J., Brown, C., Duffy, S., Raffle, A., Quinn, M., Zackrisson, S., Andersson, I., Janzon, L., Manjer, J., Garne, J., Gotzsche, P., Nielsen, M., Nielsen, M., Thomsen, J., Primdahl, S., Dyreborg, U., Andersen, JA, Ottesen, G., Graversen, H., Blichert-Toft, M., Christensen, I., Andersen, JA, Zahl, P., Andersen, JM, Maehlen, J., Zahl, P., Strand, B., Mahlen, J., Ostor, A., Rose, G., Barker, D., Taupin, D., Chambers, S., Corbett, M., Shadbolt, B., Towler, B., Irwig, L., Glasziou, P., Weller, D., Kewenter, J., Raffle, A., Alden, B., Quinn, M., Babb, P., Brett, M., Brett, J., Austoker, J., Brett, J., Bankhead, C., Henderson, B., Watson, E., Austoker, J., Barratt, A., McCaffery, K., Barratt, A., Lunde, I., Cockburn, J., De, L., Hurley, S., Clover, K., Posner, T., Vessey, M., Padgett, D., Yedidia, M., Kerner, J., Mandelblatt, J, Brodersen, J, Brodersen, J, Thorsen, H, Cockburn, J., Brodersen, J, Thorsen, H, Cockburn, J., Cullen, J., Schwartz, M., Lawrence, W., Selby, J., Mandelblatt, JS, Hobart, J., Williams, L., Moran, K., Thompson, A., Cortina, J., Tennant, A., McKenna, S., Hagell, P., Rosenbaum, P., Rasch, G., Andersen, E., Bartholomew, D., Raczek, A., Ware, J., Bjorner, J., Gandek, B., Haley, S., Aaronson, N., Apolone, G., Bech, P., Brazier, J., Bullinger, M., Sullivan, M., Gilbert, C., Brown, M., Cappelleri, J., Parpia, T., McKenna, S., 2007. Measuring the psychosocial consequences of screening. Health Qual. Life Outcomes 5, 3. https://doi.org/10.1186/1477-7525-5-3

Browne, M.W., 2001. An overview of analytic rotation in exploratory factor analysis. Multivariate Behav. Res. 36, 111–150. https://doi.org/10.1207/S15327906MBR3601_05

Cai, N., Kendler, K.S., Flint, J., 2018. Minimal phenotyping yields GWAS hits of low specificity for major depression. BioRxiv 44, 300.

Campbell, A., Knowles, S., 2007. A confirmatory factor analysis of the GHQ12 using a large Australian sample. Eur. J. Psychol. Assess. 23, 2–8. https://doi.org/10.1027/1015-5759.23.1.2

Campbell, A., Walker, J., Farrell, G., 2003. Confirmatory factor analysis of the GHQ-12: can I see that again? Aust. N. Z. J. Psychiatry 37, 475–483. https://doi.org/10.1046/j.1440-1614.2003.01208.x

Charon Gwynn, R., McQuistion, H.L., McVeigh, K.H., Garg, R.K., Frieden, T.R., Thorpe, L.E., 2008. Prevalence, Diagnosis, and Treatment of Depression and Generalized Anxiety Disorder in a Diverse Urban Community 59, 641–647.

Chen, F.F., 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. Struct. Equ. Model. A Multidiscip. J. 14, 464–504. https://doi.org/10.1080/10705510701301834

Cheung, Y.B., 2002. A confirmatory factor analysis of the 12-item General Health Questionnaire among older people 739–744.

Conway, J.M., Huffcutt, A.I., 2003. A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. Organ. Res. Methods 6, 147–168. https://doi.org/10.1177/1094428103251541

Creed, P.A., Evans, B.M., 2002. Personality, well-being and deprivation theory. Pers. Individ. Dif. 33, 1045–1054. https://doi.org/10.1016/S0191-8869(01)00210-0

Cronbach, L.J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334. https://doi.org/10.1007/BF02310555

Davies, S., 2013. Annual Report of the Chief Medical Officer 2013, Public Health Mental Priorities: Investing in the evidence 320.

Davies, S.C., 2018. Annual Report of the Chief Medical Officer, 2018: Health 2040 - Better Health Within Reach.

Dunn, T.J., Baguley, T., Brunsden, V., 2014. From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. Br. J. Psychol. 105, 399–412. https://doi.org/10.1111/bjop.12046

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., Strahan, E.J., 1999. Evaluating the use of exploratory factor analysis in psychological research. Psychol. Methods 4, 272–299. https://doi.org/10.1037//1082-989X.4.3.272

Faris, R.E.L., Dunham, H.W., 1939. Mental disorders in urban areas: an ecological study of schizophrenia and other psychoses. Univ. Chicago Press, Oxford, England.

Fernandes, H.M., Vasconcelos-Raposo, J., 2013. Factorial validity and invariance of the GHQ-12 among clinical and nonclinical samples. Assessment 20, 219–29. https://doi.org/10.1177/1073191112465768

French, D.J., Tait, R.J., 2004. Measurement invariance in the General Health Questionnaire-12 in young Australian adolescents. Eur. Child Adolesc. Psychiatry 13, 1–7. https://doi.org/10.1007/s00787-004-0345-7

Gao, F., Luo, N., Thumboo, J., Fones, C., Li, S.-C., Cheung, Y.-B., 2004. Does the 12-item General Health Questionnaire contain multiple factors and do we need them? Health Qual. Life Outcomes 2, 63. https://doi.org/10.1186/1477-7525-2-63

Garrard, J., Rolnick, S.J., Nitz, N.M., Luepke, L., Jackson, J., Fischer, L.R., Leibson, C., Bland, P.C., Heinrich, R., Waller, L.A., 1998. Clinical detection of depression among community-based elderly people with self-reported symptoms of depression. J. Gerontol. A. Biol. Sci. Med. Sci. 53, M92–M101.

Gelman, a, Gelman, a, Meng, X.-L., Meng, X.-L., Stern, H., Stern, H., 1996. Posterior predictive assessment of model fitness via realized discrepancies. Vol.6, No.4. Stat. Sin. 6, 733–807. https://doi.org/10.1.1.142.9951

Gnambs, T., Staufenbiel, T., 2018. The structure of the General Health Questionnaire (GHQ-12): two meta-analytic factor analyses. Health Psychol. Rev. 12, 179–194. https://doi.org/10.1080/17437199.2018.1426484

Goldberg, D., Bridges, K., 1987. Screening for psychiatric illness in general practice: the general practitioner

versus the screening questionnaire. J. R. Coll. Gen. Pract. 37, 15–18.

Goldberg, D., Gater, R., Sartorius, N., Ustün, T.B., Piccinelli, M., Gureje, O., Rutter, M., 1997. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychol. Med. 27, 191–197.

Goldberg, D.P., Hillier, V.F., 1979. A scaled version of the General Health Questionnaire. Psychol. Med. 9, 139–145. https://doi.org/10.1017/S0033291700021644

Goldberg, D.P., Oldehinkel, T., Ormel, J., 1998. Why GHQ threshold varies from one place to another. Psychol. Med. 28, 915–21. https://doi.org/10.1017/S0033291798006874

Gouveia, V. V, Barbosa, G.A., Oliveira Andrade, E. De, Carneiro, M.B., 2010. Factorial validity and reliability of the General Health Questionnaire (GHQ-12) in the Brazilian physician population. Cad. Saude Publica 26, 1439–1445. https://doi.org/10.1590/S0102-311X2010000700023

Graetz, B., 1991. Multidimensional properties of the general health questionnaire. Soc. Psychiatry Psychiatr. Epidemiol. 132–138.

Grotzinger, A.D., Rhemtulla, M., de Vlaming, R., Ritchie, S.J., Mallard, T.T., Hill, W.D., Ip, H.F., Marioni, R.E., McIntosh, A.M., Deary, I.J., Koellinger, P.D., Harden, K.P., Nivard, M.G., Tucker-Drob, E.M., 2019. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nat. Hum. Behav. 3, 513–525. https://doi.org/10.1038/s41562-019-0566-x

Hankins, M., 2008a. The reliability of the twelve-item general health questionnaire (GHQ-12) under realistic assumptions. BMC Public Health 8, 355. https://doi.org/10.1186/1471-2458-8-355

Hankins, M., 2008b. Clinical Practice and Epidemiology The factor structure of the twelve item General Health Questionnaire (GHQ-12): the result of negative phrasing ? 8, 1–8. https://doi.org/10.1186/1745-0179-4-Received

Hu, L., Bentler, P.M., 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct. Equ. Model. A Multidiscip. J. 6, 1–55. https://doi.org/10.1080/10705519909540118

Hu, Y., Stewart-Brown, S., Twigg, L., Weich, S., 2007. Can the 12-item General Health Questionnaire be used to measure positive mental health? Psychol. Med. 37, 1005–13. https://doi.org/10.1017/S0033291707009993

Jöreskog, K.G., 1969. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika 34, 183–202. https://doi.org/10.1007/BF02289343

Kalliath, T.J., O'Driscoll, M.P., Brough, P., 2004. A confirmatory factor analysis of the General Health Questionnaire-12. Stress Heal. 20, 11–20. https://doi.org/10.1002/smi.993

Khan, A., Mad Shah, I., Khan, F., Suhail, S., 2013. Reliability and validity assessment of 12 items general health questionnaire (GHQ: 12) among Pakistani university teachers. World Appl. Sci. J. 24, 603–608. https://doi.org/10.5829/idosi.wasj.2013.24.05.13212

Kiliç, C., Rezaki, M., Rezaki, B., Kaplan, I., Özgen, G., Sağduyu, A., Oztŕk, M.O., 1997. General Health Questionnaire (GHQ12 and GHQ28): Psychometric properties and factor structure of the scales in a Turkish primary care sample. Soc. Psychiatry Psychiatr. Epidemiol. 32, 327–331. https://doi.org/10.1007/BF00805437

Kish, L., 1954. Differentiation in Metropolitan Areas. Am. Sociol. Rev. 19, 388–398. https://doi.org/10.2307/2087457

Knies, G., 2017. Understanding Society: Wave 1-7, 2009-2016 and harmonised British Household Panel Survey: Waves 1-18, 1991-2009, User Guide., Colchester. University of Essex. https://doi.org/10.2307/3348243

Lecrubier, Y., 2007. Widespread underrecognition and undertreatment of anxiety and mood disorders: Results from 3 European studies. J. Clin. Psychiatry 68, 36–41.

Lynn, P., Knies, G., 2016. Understanding Society: The UK Household Longitudinal Study Waves 1-5 Quality Profile. Inst. Soc. ….

Marsh, H.W., Bailey, M., 1991. Confirmatory Factor Analyses of Multitrait-Multimethod Data: A Comparison of Alternative Models. Appl. Psychol. Meas. 15, 47–70. https://doi.org/10.1177/014662169101500106

Marsh, H.W., Hau, K.-T., Wen, Z., 2004. In Search of Golden Rule : Comment on Hypothesis Testing Approaches to Setting Cutoff Value for Fit Indexes and Danger in Overgeneralizing Hu and Bentler's (1999) Finding. Struct. Equ. Model. 11, 320–341. https://doi.org/10.1207/s15328007sem1103

Marsh, H.W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A.J.S., Trautwein, U., Nagengast, B., 2010a. A new look at the big five factor structure through exploratory structural equation modeling. Psychol. Assess. 22, 471–91. https://doi.org/10.1037/a0019227

Marsh, H.W., Lüdtke, O., Muthén, B., Asparouhov, T., Morin, A.J.S., Trautwein, U., Nagengast, B., 2010b. A new look at the big five factor structure through exploratory structural equation modeling. Psychol. Assess. 22, 471–491. https://doi.org/10.1037/a0019227

Marsh, H.W., Morin, A.J.S., Parker, P.D., Kaur, G., 2014. Exploratory structural equation modeling: an integration of the best features of exploratory and confirmatory factor analysis. Annu. Rev. Clin. Psychol. 10, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Marsh, H.W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A.J.S., Trautwein, U., 2009. Exploratory Structural Equation Modeling, Integrating CFA and EFA: Application to Students' Evaluations of University Teaching, Structural Equation Modeling: A Multidisciplinary Journal. https://doi.org/10.1080/10705510903008220

Martin, A.J., 1999. Assessing the Multidimensionality of the 12-Item General Health Questionnaire. Psychol. Rep. 84, 927–935.

Martin, C.R., Newell, R.J., 2005. The factor structure of the 12-item General Health Questionnaire in individuals with facial disfigurement. J. Psychosom. Res. 59, 193–9. https://doi.org/10.1016/j.jpsychores.2005.02.020

McDonald, R.P., 2005. Semiconfirmatory Factor Analysis: The Example of Anxiety and Depression. Struct. Equ. Model. A Multidiscip. J. 12, 163–172. https://doi.org/10.1207/s15328007sem1201_9

McFall, S.L., 2011. Understanding Society - The UK household longitudinal study: Wave 1, 2009-2010, User Manual. Colchester.

Mindell, J., Biddulph, J.P., Hirani, V., Stamatakis, E., Craig, R., Nunn, S., Shelton, N., 2012. Cohort profile: The health survey for england. Int. J. Epidemiol. 41, 1585–1593. https://doi.org/10.1093/ije/dyr199

Mukuria, C., Rowen, D., Peasgood, T., Brazier, J., 2014. An empirical comparison of well-being measures used in UK 1–55.

Muthén, B., Asparouhov, T., 2012. Bayesian structural equation modeling: a more flexible representation of substantive theory. Psychol. Methods 17, 313–35. https://doi.org/10.1037/a0026802

Muthén, B., Asparouhov, T., 2010. Bayesian SEM : A more flexible representation of substantive theory.

Muthen, L., Muthen, B., 2017. Mplus User's Guide. https://doi.org/10.1111/j.1600-0447.2011.01711.x

Oberski, D.L., 2014. Evaluating measurement invariance in categorical data latent variable models with the EPC -interest Evaluating measurement invariance in categorical data latent variable models with the EPC -

interest.

Padrón, A., Galán, I., Durbán, M., Gandarillas, A., Rodríguez-Artalejo, F., 2012. Confirmatory factor analysis of the General Health Questionnaire (GHQ-12) in Spanish adolescents. Qual. Life Res. 21, 1291–8. https://doi.org/10.1007/s11136-011-0038-x

Patalay, P., Fitzsimons, E., 2016. Correlates of Mental Illness and Wellbeing in Children : 55. https://doi.org/10.1016/j.jaac.2016.05.019

Penninkilampi-Kerola, V., Miettunen, J., Ebeling, H., 2006. A comparative assessment of the factor structures and psychometric properties of the GHQ-12 and the GHQ-20 based on data from a Finnish population-based sample. Scand. J. Psychol. 47, 431–40. https://doi.org/10.1111/j.1467-9450.2006.00551.x

Pevalin, D.J., 2000. Multiple applications of the GHQ-12 in a general population sample: an investigation of long-term retest effects. Soc. Psychiatry Psychiatr. Epidemiol. 35, 508–512. https://doi.org/10.1007/s001270050272

Picardi, a, Abeni, D., Pasquini, P., 2001. Assessing psychological distress in patients with skin diseases: reliability, validity and factor structure of the GHQ-12. J. Eur. Acad. Dermatology Venereol. 15, 410–417. https://doi.org/10.1046/j.1468-3083.2001.00336.x

Politi, P.L., Piccinelli, M., Wilkinson, G., 1994. Reliability, validity and factor structure of the 12-item General Health Questionnaire among young males in Italy. Acta Psychiatr. Scand. 90, 432–437. https://doi.org/10.1111/j.1600-0447.1994.tb01620.x

Propper, C., Jones, K., Bolster, A., Burgess, S., Johnston, R., Sarker, R., 2005. Local neighbourhood and mental health: evidence from the UK. Soc. Sci. Med. 61, 2065–83. https://doi.org/10.1016/j.socscimed.2005.04.013

Romppel, M., Braehler, E., Roth, M., Glaesmer, H., 2013. What is the General Health Questionnaire-12 assessing?: Dimensionality and psychometric properties of the General Health Questionnaire-12 in a large scale German population sample. Compr. Psychiatry 54, 406–413. https://doi.org/10.1016/j.comppsych.2012.10.010

Sánchez-López, M.D.P., Dresch, V., 2008. The 12-Item General Health Questionnaire (GHQ-12): reliability, external validity and factor structure in the Spanish population. Psicothema 20, 839–43.

Schmitt, T. a., 2011. Current Methodological Considerations in Exploratory and Confirmatory Factor Analysis. J. Psychoeduc. Assess. 29, 304–321. https://doi.org/10.1177/0734282911406653

Schmitz, N., Kruse, J., Tress, W., 1999. Psychometric properties of the General Health Questionnaire (GHQ-12) in a German primary care sample. Acta Psychiatr. Scand. 100, 462–468. https://doi.org/10.1111/j.1600-0447.1999.tb10898.x

Seddig, D., 2018. Approximate measurement invariance and longitudinal confirmatory factor analysis : concept and application with panel data 12, 29–41.

Shevlin, M., Adamson, G., 2005. Alternative factor models and factorial invariance of the GHQ-12: a large sample analysis using confirmatory factor analysis. Psychol. Assess. 17, 231–236. https://doi.org/10.1037/1040-3590.17.2.231

Smith, A.B., Fallowfield, L.J., Stark, D.P., Velikova, G., Jenkins, V., 2010. A Rasch and confirmatory factor analysis of the general health questionnaire (GHQ)--12. Health Qual. Life Outcomes 8, 45. https://doi.org/10.1186/1477-7525-8-45

Smith, A.B., Oluboyede, Y., West, R., Hewison, J., House, A.O., 2013. The factor structure of the GHQ-12: the interaction between item phrasing, variance and levels of distress. Qual. Life Res. 22, 145–52.

https://doi.org/10.1007/s11136-012-0133-7

Spiegelhalter, D., Best, N., Carlin, B., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. … 583–639.

Stewart-Brown, S., Tennant, A., Tennant, R., Platt, S., Parkinson, J., Weich, S., 2009. Internal construct validity of the Warwick-Edinburgh Mental Well-being Scale (WEMWBS): a Rasch analysis using data from the Scottish Health Education Population Survey. Health Qual. Life Outcomes 7, 15. https://doi.org/10.1186/1477-7525-7-15

Stochl, J., Böhnke, J.R., Pickett, K.E., Croudace, T.J., 2016. An evaluation of computerized adaptive testing for general psychological distress: Combining GHQ-12 and Affectometer-2 in an item bank for public mental health research. BMC Med. Res. Methodol. 16, 1–15. https://doi.org/10.1186/s12874-016-0158-7

Suzuki, H., Kaneita, Y., Osaki, Y., Minowa, M., Kanda, H., Suzuki, K., Wada, K., Hayashi, K., Tanihata, T., Ohida, T., 2011. Clarification of the factor structure of the 12-item General Health Questionnaire among Japanese adolescents and associated sleep status. Psychiatry Res. 188, 138–46. https://doi.org/10.1016/j.psychres.2010.10.025

Sweeting, H., Young, R., West, P., 2009. GHQ increases among Scottish 15 year olds 1987-2006. Soc. Psychiatry Psychiatr. Epidemiol. 44, 579–86. https://doi.org/10.1007/s00127-008-0462-6

Tait, R.J., French, D.J., Hulse, G.K., 2003. Validity and psychometric properties of the General Health Questionnaire-12 in young Australian adolescents. Aust. N. Z. J. Psychiatry 37, 374–381. https://doi.org/10.1046/j.1440-1614.2003.01133.x

Tennant, R., Hiller, L., Fishwick, R., Platt, S., Joseph, S., Weich, S., Parkinson, J., Secker, J., Stewart-Brown, S., 2007. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): development and UK validation. Health Qual. Life Outcomes 5, 63. https://doi.org/10.1186/1477-7525-5-63

Thomson, R.M., Katikireddi, S.V., 2018. Mental health and the jilted generation: Using age-period-cohort analysis to assess differential trends in young people's mental health following the Great Recession and austerity in England. Soc. Sci. Med. 214, 133–143. https://doi.org/10.1016/j.socscimed.2018.08.034

Thomson, R.M., Vittal, S., 2018. Social Science & Medicine Mental health and the jilted generation: Using age-period-cohort analysis to assess differential trends in young people ' s mental health following the Great Recession and austerity in England. Soc. Sci. Med. 214, 133–143. https://doi.org/10.1016/j.socscimed.2018.08.034

Vanheule, S., Bogaerts, S., 2005. The factorial structure of the GHQ-12. Stress Heal. 21, 217–222. https://doi.org/10.1002/smi.1058

Verheij, R., 1996. Explaining urban-rural variations in health: a review of interactions between individual and environment. Soc. Sci. Med. 42, 923–935.

Wang, L., Lin, W., 2011. Wording effects and the dimensionality of the General Health Questionnaire (GHQ-12). Pers. Individ. Dif. 50, 1056–1061. https://doi.org/10.1016/j.paid.2011.01.024

Weich, S., Holt, G., Twigg, L., Jones, K., Lewis, G., 2003. Geographic Variation in the Prevalence of Common Mental Disorders in Britain: A Multilevel Investigation. Am. J. Epidemiol. 157, 730–737. https://doi.org/10.1093/aje/kwg035

Werneke, U., Goldberg, D.P., Yalcin, I., Üstün, B.T., 2000. The stability of the factor structure of the General Health Questionnaire. Psychol. Med. 30, 823–829. https://doi.org/10.1017/S0033291799002287

Westerhof, G.J., Keyes, Æ.C.L.M., 2010. Mental Illness and Mental Health : The Two Continua Model Across the Lifespan 110–119. https://doi.org/10.1007/s10804-009-9082-y

Winefield, H.R., Goldney, R.D., Winefield, A.H., Tiggemann, M., 1989. The General Health Questionnaire: Reliability and validity for Australian youth. Aust. N. Z. J. Psychiatry 23, 53–58. https://doi.org/10.3109/00048678909062592

World Health Organization, 2013. Mental Health Action Plan 2013-2020. WHO Libr. Cat. DataLibrary Cat. Data 1–44. https://doi.org/ISBN 978 92 4 150602 1

Worsley, A., Gribbin, C.C., 1977. A factor analytic study on the twelve item general health questionnaire. Aust. N. Z. J. Psychiatry 11, 260–72. https://doi.org/10.3109/00048677709159577

Ye, S., 2009. Factor structure of the General Health Questionnaire (GHQ-12): The role of wording effects. Pers. Individ. Dif. 46, 197–201. https://doi.org/10.1016/j.paid.2008.09.027

Demographic Characteristic	**Full Respondents** *(N=39700)*	**Full and Partial Respondents** *(N=40452)*
Age	Mean = 45.8, S.D. = 18.0	Mean = 45.8, S.D. = 18.0
Sex		
Male	43.9%	43.8%
Female	56.1%	56.2%
Ethnicity		
White	84.6%	83.1%
Black	4.3%	4.6%
Asian	7.9%	9.1%
Mixed	3.2%	3.3%
Job Classification		
Not in Employment	43.5%	43.5%
Professional	3.6%	3.6%
Managerial/Technical	21.0%	20.9%
Skilled Non-Manual	12.7%	12.6%
Skilled Manual	9.0%	9.0%
Partly Skilled	8.2%	8.2%
Unskilled	1.9%	1.9%
Highest Educational Qualification		
Higher Education	34.0%	33.8%
A-Level or equivalent	19.3%	19.1%
GCSE or equivalent	20.9%	20.9%
Other Qualification	4.9%	5.0%
No Qualifications	20.9%	21.2%

Table 1: Demographic characteristics of full and partial respondents to the GHQ-12 in Wave 1 of Understanding Society.

EFA	*WLSMV*				*Bayesian*		
Number of Factors	CFI	TLI	RMSEA	SRMR	Posterior P-Value	2.5% CI	97.5% CI
2	0.976	0.964	0.080	0.036	0.000	2668.266	2993.799
3	0.993	0.987	0.048	0.017	0.000	457.804	620.840
4	0.997	0.992	0.037	0.011	0.000	197.311	331.851
5	0.999	0.995	0.029	0.007	0.000	55.562	1073.603

Table 1: Bayesian and ML Fit statistics for EFA factor solutions for the GHQ-12 with 2, 3, 4 and 5 factors. For further information on fit statistics see Supplementary Material 2.

GHQ-12 Items	F1 Lowered Self Worth	F2 Social Dysfunction	F3 Stress	F4 Emotional Coping
1. Able to Concentrate		0.659 (0.008)	0.227 (0.009)	
2. Loss of Sleep	0.632 (0.014)		0.427 (0.007)	
3. Playing a useful part		0.646 (0.007)		
4. Capable of decisions		0.907 (0.013)		-0.322 (0.017)
5. Constantly under strain	0.753 (0.018)		0.617 (0.008)	
6. Problem overcoming difficulties	0.754 (0.011)		0.330 (0.006)	
7. Enjoy day-to-day activities		0.691 (0.009)	0.200 (0.008)	0.244 (0.010)
8. Ability to face problems		0.746 (0.006)		
9. Feel unhappy/depressed	0.746 (0.009)		0.261 (0.006)	0.216 (0.009)
10. Losing confidence	0.909 (0.006)			
11. Think of self as worthless	0.887 (0.007)			
12. Feeling reasonably happy		0.577 (0.010)		0.295 (0.011)

Table 2: Standardised Factor Loadings for 4-Factor ESEM Model of the GHQ-12 using Bayesian estimation. Standard Errors in parentheses.

FACTOR CORRELATIONS	F1 Lowered Self-Worth	F2 Social Dysfunction	F3 Stress	F4 Emotional Coping
F1. Lowered Self Worth	1.000			
F2. Social Dysfunction	0.680	1.000		
F3. Stress	0.178	0.111	1.000	
F4. Emotional Coping	0.411	0.487	0.271	1.000

Table 3: Modelled Factor Correlations from the Four Factor ESEM Solution for the GHQ-12

Authors	*Date*	*N*	*Factor Structure Details*
Initial GHQ-12 Goldberg Formulation	1972	200	1 Factor – Baseline unidimensional model, all items specified to load on a single factor.
Hankins	2008	3705	1 Factor – Unidimensional but with error covariance specified on negatively phrased items.
Andrich & Van Schoenbroeck	1989	491	2 Factor – Split into positive and negative items, each constituting a separate dimension.
Kilic et al.	1997	1307	2 Factor – Anxiety/Depression, Social Dysfunction
Worsley and Gribbin	1977	603	3 Factor – Anhedonia-Sleep Disturbance, Social Performance, Loss of Confidence, specifies cross loadings on items 1 (F1, F2), 6 (F1, F3), 9 (F1, F3) and 12 (F1, F2)
Graetz	1991	8998	3 Factor – Anxiety/Depression, Social Dysfunction, Loss of Confidence
Sanchez-Lopez & Dresch	2008	1001	3 Factor – Successful Coping, Self-Esteem, Stress, Includes one non-zero cross-loading on Item 9 for F2 and F3.

Table 4: Seven exemplar studies of the range of factor analytical structures obtained from the GHQ-12 data. See Supplementary Materials for diagrammatic representation of structures.

Factor Structure	Factor No.	SRMR	CFI	TLI	RMSEA	Posterior predictive P	Lower Bound	Upper Bound	Mean Absolute Factor Correlation
Unidim-ensional	1	0.062	0.919	0.901	0.131	0.000	7149.856	7670.342	-
Hankins	1	0.029	0.983	0.972	0.070	0.000	1505.348	1764.743	-
Kilic et al.	2	0.053	0.937	0.922	0.123	0.000	2665.799	3000.182	0.822
Andrich & van Schoubroeck	2	0.036	0.973	0.966	0.077	0.000	2653.599	2983.876	0.729
Sanchez-Lopez & Dresch	3	0.033	0.976	0.969	0.074	0.000	2824.202	3152.636	0.772
Graetz	3	0.032	0.978	0.972	0.070	0.000	2242.227	2552.853	0.768
Worsley & Gribbin	3	0.021	0.990	0.985	0.050	0.000	1199.430	1430.393	0.665
ESEM 4Fac	4	0.009	0.997	0.992	0.037	0.000	376.571	526.539	0.356

Table 6: Bayesian and WLSMV Fit Statistics for the proposed ESEM factor structure alongside the 7 defined in Table 5.

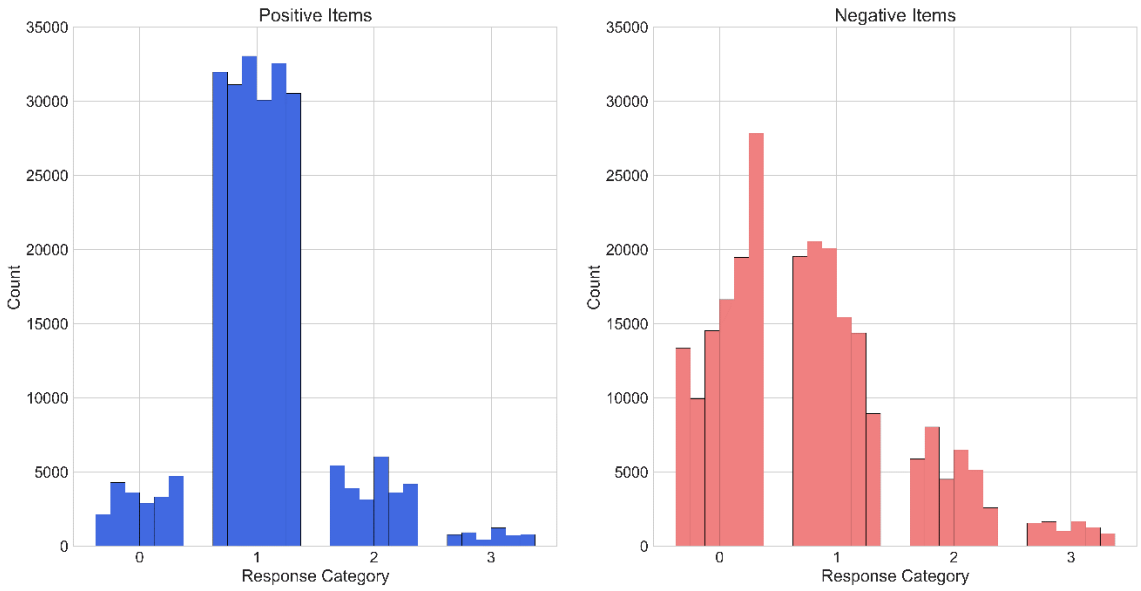

Figure 1: Graph illustrating the differential response patterning of positively (1,3,4,7,8,12) and negatively (2,5,6,9,10,11) phrased items in the GHQ-12 from Wave 1 of Understanding Society.